Multi-Stage Multi-Modal Pre-Training For Automatic Speech Recognition

Yash Jain[†§1], **David Chan**[*§1], **Pranav Dheram**[§], **Aparna Khare**[§], **Olabanji Shonibare**[§], **Venkatesh Ravichandran**[§], **Shalini Ghosh**[§]
[†]Georgia Institute of Technology
[*]University of California, Berkeley (EECS)
[§]Amazon Alexa AI
yashjain@gatech.edu

Abstract

Recent advances in machine learning have demonstrated that multi-modal pre-training can improve automatic speech recognition (ASR) performance compared to randomly initialized models , even when models are fine-tuned on uni-modal tasks. Existing multi-modal pre-training methods for the ASR task have primarily focused on single-stage pre-training where a single unsupervised task is used for pre-training followed by fine-tuning on the downstream task. In this work, we introduce a novel method combining multi-modal and multi-task unsupervised pre-training with a translation-based supervised mid-training approach. We empirically demonstrate that such a multi-stage approach leads to relative word error rate (WER) improvements of up to 38.45% over baselines on both Librispeech and SUPERB. Additionally, we share several important findings for choosing pre-training methods and datasets.

Keywords: Multi-modal, Speech Recognition, Self-supervised, Pre-training, Mid-training

1. Introduction

Despite progress in large-scale pre-training for automatic speech recognition (ASR) (Chen et al., 2022; Hsu and Shi, 2022; Chan et al., 2022), uni-modal (speech-only) ASR remains a challenging task, particularly when faced with rare words and noisy acoustic conditions. When understanding spoken phonemes, the model must correctly discern both speaker-specific patterns (e.g., accent, prosody) and global noise patterns (e.g., background noise, intermittent interruptions, confounding speakers). Recent work in natural language processing (NLP) (Tu et al., 2020; Hoffmann et al., 2022), robotics (Mandlekar et al., 2022; Kuhar et al., 2023; Khazatsky et al., 2024) and computer vision (Goyal et al., 2022; Ramanujan et al., 2023; Jain et al., 2024) has demonstrated that exposing models to a *high diversity* of data during pre-training is essential in building robust representations.

Similarly, recent works in the ASR community have corroborated these results. Shi et al. (2022) and Hsu and Shi (2022) demonstrated that pre-training on large-scale audio-visual data (or audio-only data), in the form of lip-reading videos, leads to better performance on the lip-reading task. Chan et al. (2022) showed that exposing models to video data during pre-training led to performance improvements not only when visual input is available at training time, but also when *only audio is available at test time*.

Chan et al. (2022) also demonstrated that adding visual information from non-speech specific videos (leveraging the Kinetics dataset (Carreira and Zisserman, 2017)) is only a small portion of the possible augmentations that can be made during pre-training. In this work, we not only explore two new audio-visual pre-training sources, but also leverage a translation task with English speech input as a new mid-training task to consolidate information learned during the pre-training phase. Further, while Chan et al. (2022) explore an attention-based transfer-learning framework based on k-means clustering for pre-training, we simplify the pre-training architecture significantly, and explore several pre-training objectives beyond masked cluster prediction. Our primary contributions are as follows:

1. We perform large-scale evaluation of multiple audio-visual pre-training methods (MAE, CLR) using several pre-training datasets (Kinetics, VoxCeleb2, LRS3) with varying characteristics. We evaluate them on the ASR task and the SUPERB benchmark, showing how multi-modal pre-training is affected by key dataset characteristics.

2. We show that pre-training with audio-visual data, particularly data from speech-specific audio-visual datasets can improve word error rate (WER) up to 30.8% relative compared to randomly initialized baseline models on speech-only test data.

3. We introduce a novel mid-training stage between the pre-training and fine-tuning steps, using speech translation as the mid-training task. The mid-training stage improves WER by 38.45% relative on the Librispeech test-clean

[1]Work done during an internship at Amazon Alexa AI.

LREC-COLING 2024, pages 11969–11980
20-25 May, 2024.

dataset, and by 26.18% relative on the test-other dataset compared to audio-visual pre-training only baseline. The technique also shows improvements on several tasks (Keyword Spotting, Intent Classification, Phoneme Recognition, and Speaker Diarization) in the SUPERB (Yang et al., 2021) benchmark.

2. Background

Representation learning methods like Contrastive Predictive Coding (Oord et al., 2018) and Wav2Vec (Schneider et al., 2019) have shown significant promise when applied to ASR. Methods for large-scale pre-training for ASR can be categorized into two methods: masked autoencoding methods (Hsu et al., 2021; Chen et al., 2022), and contrastive learning (Baevski et al., 2020). While traditionally self-supervised methods are trained on a single target loss, other methods have been proposed which leverage multiple pre-training targets. Pascual et al. (2019); Talnikar et al. (2021); Wang et al. (2021a) all optimize a combination of uni-modal supervised losses and recently, approaches such as W2v-BERT (Chung et al., 2021) and JUST (Bai et al., 2022) have combined contrastive approaches with masked auto-encoding to build robust self-supervised speech representations. Similarly, while most self-supervised methods are pre-trained on a single dataset, Radford et al. (2022); Narayanan et al. (2018); Likhomanenko et al. (2020); Chan et al. (2021) have all demonstrated that a wide mix of data is essential for pre-training. In this work, we target both of these problems: use a combination of losses, and pre-training stages under different datasets to improve the learned multi-modal representations.

Audio-visual data provides diverse information for representation learning. Shi et al. (2022) demonstrate improvements on ASR when visual input is available (at both training and test time), and methods such as u-HuBERT (Hsu and Shi, 2022) extend such pre-training approaches to applications where both uni-modal and multi-modal data are available at training-time (but still require multi-modal data for inference). Later work by Chan et al. (2022) demonstrated that pre-training with paired audio-visual data, can even improve performance on *uni-modal* datasets.

In addition to multiple modalities, pre-training with multiple languages has also been explored in the literature. Radford et al. (2022) demonstrate that pre-training with a wide range of inputs from several languages improves ASR performance across all of the studied languages. (Lahiri et al., 2021) show that leveraging self-supervised learning (SSL) for knowledge transfer across languages can yield WER improvements of up to 3.55% relative WER on target languages, and Karimi et al. (2022) demonstrate that in almost all cases, even out-of-domain multi-lingual data can improve WER in single and multi-speaker conversations and dictation tasks.

3. Methods

Our method (Figure 1), consists of a multi-stage multi-modal pre-training approach, followed by a fine-tuning stage on downstream tasks. We describe our method in this section.

3.1. Pre-Training Tasks

We experiment with two pre-training strategies that differ in the granularity of information they extract. The first method, Masked Autoencoder (MAE), learns local features by reconstructing masked parts of speech and video. The second method, Contrastive Learning (CLR), focuses on global features by using pooled audio-visual features from the same video as positive pairs while other combinations of audio-visual pairs as negatives. The two pre-training strategies help us compare the effects of local and global feature learning against the visual-audio dataset characteristics, for *eg.*, Kinetics dataset (Carreira et al., 2018) has non-speech audio streams, while LRS-3 (Afouras et al., 2018) and Voxceleb2 (Chung et al., 2018) datasets have videos with speech.

Masked Autoencoding (MAE): Traditional MAE approaches for ASR pre-training have focused on token-based reconstruction (Hsu et al., 2021; Shi et al., 2022; Chan et al., 2022). However these methods have the drawback of requiring a separate quantization method, which can add significant training complexity. We simplify the encoder to directly reconstruct features from the original masked signal.

Our MAE approach consists of three encoders: $\mathcal{E}_a$, a masked audio-specific encoder based on the encoder in Chen et al. (2022), $\mathcal{E}_v$, a masked video-specific encoder based on Tong et al. (2022), and $\mathcal{D}_{a+v}$, a joint transformer decoder with the same structure as in Devlin et al. (2018).

Let $a \in \mathbb{R}^{T_a \times F}$ be the set of audio input frames (we use f-dimensional log-filterbank energies (LFBE)), and $v \in \mathbb{R}^{H//P_H \times W//P_W \times T_v//P_T \times (P_H P_W P_T)}$ be a set of video frames, which have been subdivided into (P_H, P_W, P_T) voxels. T_a refers to number of audio frames and T_v are number of video frames of height H and width W. To generate the input sequence to $\mathcal{E}_a$, we randomly mask a fraction ϕ of the audio frames with 0s (masking), and generate the embedded audio $e_a = \mathcal{E}_a(a)$. We use a similar process to mask voxels, to generate $e_v = \mathcal{E}_v(v)$.

The encoded representations e_a and e_v are passed through the common decoder $\mathcal{D}_{a+v}$ to pro-

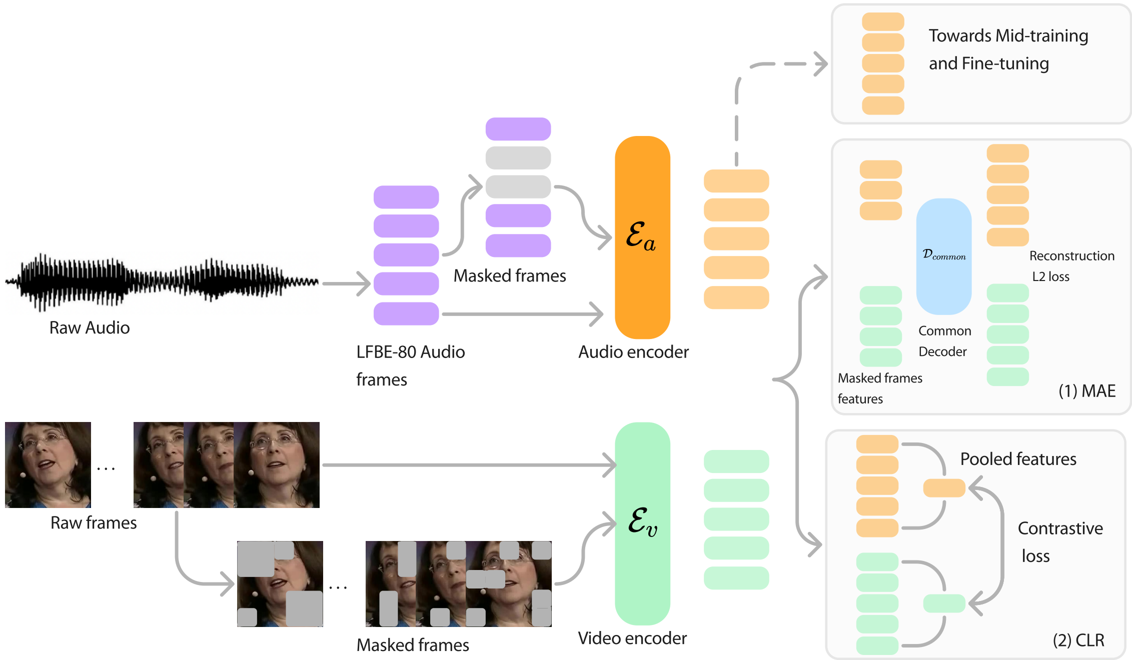

Figure 1: Overview of multi-modal training strategy. Raw audio and video features are extracted from source data. These features are then passed through the audio and video encoders to get features which are further processed as (1) MAE: the masked encoded features are reconstructed through a common decoder successively and are compared against original input using L2 loss, (2) CLR: contrastive learning applied to spatio-temporally pooled audio and video encoded features, and (3) the trained audio encoder is further used for mid-training (translation task) and then for downstream tasks.

duce $o_a = \mathcal{D}_{a+v}(e_a)$ and $o_v = \mathcal{D}_{a+v}(e_v)$ respectively. The common decoder $\mathcal{D}_{a+v}$ ensures that the representations e_a and e_v are projected to the same representation space. The final MAE loss is computed as the squared L2 distance between o_a and a, and o_v and v:

$$\mathcal{L}_{\text{MAE}} = ||\mathcal{D}_{a+v}(\mathcal{E}_a(a)) - a||_2^2 + ||\mathcal{D}_{a+v}(\mathcal{E}_v(v)) - v||_2^2 \quad (1)$$

Contrastive Learning (CLR): Contrastive Learning aims to learn representations using a contrastive loss that minimizes the distance between similar points and maximizes the distance between dissimilar points in a latent space. For contrastive learning, following Radford et al. (2021) and Xu et al. (2021), we use the modality specific encodings e_a and e_v to generate $a^{enc} = \text{Pool}(e_a)$, where the pooling operation is a temporal average, and $v^{enc} = \text{Pool}(e_v)$, where the pooling operation is a spatio-temporal average. While other pooling operations like attention pooling are possible, we found that the spatio-temporal average captures consistent low-frequency global information, which correlates well with the information shared with the visual modality (unlike high-frequency information, which is often not evident from the visual modality). The self-supervised contrastive loss for a batch of samples $a_i^{enc}, 1 \le i \le N$, and $v_i^{enc}, 1 \le i \le N$ is computed as

$$\mathcal{L}^i_{\text{contrastive}} = -\log\left(\frac{\exp(a_i^{enc} \cdot v_i^{enc})}{\sum_{k=1}^{N} \mathbb{1}_{[k \neq i]} \exp(a_i^{enc} \cdot v_k^{enc})}\right) \quad (2)$$

$$\mathcal{L}_{\text{CLR}} = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}^i_{\text{contrastive}} \quad (3)$$

MAE + CLR: In this setup, we combine the benefits of learning local features using MAE with learning global features using CLR as shown in Figure 1. Both pre-training losses are added with equal weights, similar to Chung et al. (2021) to compute the final loss as

$$\mathcal{L}_{\text{MAE+CLR}} = \frac{\mathcal{L}_{\text{MAE}} + \mathcal{L}_{\text{CLR}}}{2} \quad (4)$$

Pre-training Datasets: We use three datasets for pre-training. The Kinetics-600 dataset (Carreira et al., 2018) has 966 hours of audio-visual data for activity recognition, with a focus on the environment or instrument used. The videos contains non-speech audio data and have been used previously for audio-visual training (Chan et al., 2022). Voxceleb2 (Chung et al., 2018) provides 2380 hours of multi-lingual speaker recognition data with challenging acoustics and comprehensive lip and facial movements. LRS3 (Afouras et al., 2018) features 346 hours of clean, multi-modal spoken sentence data from TED and TEDx videos. The speech data in Voxceleb2 is has noisy acoustic conditions whereas LRS-3 has clean speech with speakers talking to a close-talk microphone. These datasets allow for exploring the impact of clean-speech/noisy-speech/non-speech videos and pre-training techniques on the ASR task (subsection 3.3).

3.2. Mid-Training: Speech Translation

To improve performance of the pre-trained audio-visual models on the downstream tasks, we introduce a mid-training task that bridges the gap between pre-training and fine-tuning. Our approach transfers the learned distribution of the pre-trained model towards the distribution required for the downstream task, while discarding irrelevant information.

The mid-training task is designed to provide a low-cost warm-up for the pre-trained model, which can accurately represent various characteristics of the data. We chose to mid-train our audio encoder on the speech translation task using the MuST-C dataset (Di Gangi et al., 2019) in three languages, German, Italian and Dutch. This stage is useful for aligning the learned speech representations with the text modality which is beneficial for ASR, as shown in recent work in the speech representation learning space(Zhang et al., 2023). Our audio encoder was mid-trained until convergence on the speech translation task. This mid-training approach is the key to strong performance in downstream tasks, which we demonstrate in detail in section 4.

Using translation as a mid-training task is only one possible instantiation of the mid-training approach. In addition to translation, future work can explore other speech-centric tasks like speaker identification, implied by (Chan and Ghosh, 2022)), speaker/source separation, text to speech, and others. While we found that translation is effective in this work, we expect that each additional task will impact the downstream training process in unique ways.

3.3. Fine-Tuning

We evaluated our models by testing their performance on several downstream tasks. The fine-tuning task is distinct from the pre-training task of masked reconstruction (MAE) or contrastive learning (CLR), and the mid-training task designed to bridge the gap. Primarily, we evaluate the performance of the models on the test-clean and test-other Librispeech (Panayotov et al., 2015) datasets for ASR, as well as four tasks from the SUPERB (Yang et al., 2021) benchmark: Intent Classification (IC), Keyword Spotting (KS), Phoneme Recognition (PR) and Speaker Diarization (SD). Because our aim was to evaluate how both the pre-training and mid-training data distributions impact the final learned representations, we freeze the encoder weights during task specific fine-tuning, and fine-tune only the task specific decoder using the LS-960 dataset (for ASR) following Baevski et al. (2020) or the default datasets specified in the SUPERB benchmark (Yang et al., 2021).

3.4. Model Details

In this section, we discuss the implementation details of the different training setups across the three datasets.

Video Data Pre-processing: Videos are first resized to a resolution of 224×224 pixels, with a temporal stride of 4 and 16 frames sampled temporally. We apply random resized cropping with scale from 0.5 to 1, and random horizontal flipping following standard computer vision techniques for visual data augmentation.

Video Encoder: Our video encoding approach is similar to that of (Feichtenhofer et al., 2022). Firstly, we divide the video into a regular grid of space-time patches of dimensions $16 \times 16 \times 2$ in the (H,W,T) direction, respectively. These patches are then flattened and augmented with spatio-temporal positional embeddings (Vaswani et al., 2017).

For the Masked Autoencoder, we randomly select 60% of the patches for masking, and mask patches without replacement, while keeping the selection agnostic in the space-time domain. The remaining patches are then passed through 12 ViT encoder blocks (Dosovitskiy et al., 2020) with a hidden dimension of 768. We obtain the video encoded features of the remaining spatio-temporal patches, which are later reconstructed using a common decoder.

For Contrastive Learning, we reduce the spatial patches to a single embedding for each frame (Xu et al., 2021; Radford et al., 2021). The reduced patches are passed through a video encoder with 12 ViT encoder blocks (Dosovitskiy et al., 2020) with a hidden dimension of 768. The encoded embeddings are temporally pooled following (Xu et al., 2021), resulting in single-vector video features which can be contrasted against corresponding audio embeddings.

Audio Data Pre-processing: The audio input is re-sampled to a frequency of 16kHz. Subsequently, 80-dimensional Log-Filterbank Energy (LFBE) features are computed from the resulting audio frames. To ensure consistency in feature size, we selected the first 1000 LFBE frames for downstream processing. The frames are further sub-sampled using a 1D convolutional layer, reducing the number of audio frames to 250, following the approach of Gulati et al. (2020).

Audio Encoder: We use positional embeddings in the sub-sampled audio frames similar to video encoding, as proposed by Vaswani et al. (2017). In the Masked Autoencoder, a random mask without replacement is applied to 60% of the frames, with the visual and audio modalities sharing the same masking ratio to maintain balance in the amount of information across both modalities. The remaining frames are encoded by a Conformer (Gulati et al., 2020) with 16 layers, 4 heads, and a depth-wise convolutional kernel of size 31. Audio features are then up-sampled by a linear layer and normalized for reconstruction.

In Contrastive Learning, the sub-sampled frames are directly featurized by the Conformer blocks without any masking involved. The audio features are then temporally pooled to obtain a single feature for the audio clip, which is up-sampled and normalized. For both the Mid-training and Fine-tuning tasks, the feature output from Conformer blocks is used as input to task-specific decoders. The weights of the convolutional sub-sampling layer and Conformer blocks are the only components re-used from the pre-training stage for further steps.

Common Decoder: The Masked Autoencoder pre-training step uses a relatively small vanilla ViT (Dosovitskiy et al., 2020) decoder of hidden dimension size of 512 and 4 ViT blocks. The decoder processes a combination of the encoded and masked patches and outputs the original reconstructed signal. A shared decoder is used to sequentially reconstruct each patch.

4. Results, Analysis & Limitations

Our main results on the Librispeech dataset are shown in Table 1 and Figure 3, and demonstrate several interesting learnings:

Audio-visual Pre-training is Effective: Table 1 shows that on average in all cases, audio-visual pre-training is effective. Averaging the performance across all methods results in 6.34 $\pm$ 0.94 WER for test-clean, and 12.18 $\pm$ 0.98 for test-other. Under the null hypothesis that audio-visual pre-training is ineffective, we find significant improvements ($p = 0.035$) over the baseline.

Mid-training with all translation pairs improve ASR performance: Table 1 shows that the mid-training approach leads to significant ($p < 0.01$) improvements over pre-trained models alone, leading to relative WER improvements of 8.59%/6.77% (test-clean/test-other) with English-German pair, 18.55%/10.28% for English-Italian pair, and 13.11%/7.71% for English-Dutch pair. Surprisingly, Italian is the most effective, suggesting that choosing languages which are complementary to English may be more useful than languages which are closer to the target downstream language (English, Dutch and German all have Germanic roots, while Italian has Latin roots - see Tyshchenko et al. (2000) for a discussion on linguistic distance).

We leave it to future work to explore languages that retain very little shared information, such as Russian or Chinese. The relative performance improvements with mid-training are shown in Figure 3. The figure shows several effects which we discuss in the following sections: the model pre-trained on Kinetics dataset is most improved with mid-training, English-Italian translation is the best mid-training pair, and the model pre-trained with CLR benefits the most with mid-training.

How do pre-training datasets impact performance (Is dataset size the only factor)? Despite differences in pre-training dataset sizes, it is interesting to understand how the input mix of data impacts the overall performance of the model. Without mid-training, models pre-trained on LRS-3, the smalleset dataset, outperform all other models (6.19%/11.64% WER) on the test-other dataset. LRS-3 is a small fraction of the size of the VoxCeleb2 dataset, suggesting that the distributional makeup of the multi-modal dataset is key to pre-training performance, and *dataset size is not all that matters*. VoxCeleb2 (6.16%/12.01% WER) outperforms LRS-3 slightly on the test-clean dataset. Kinetics trails both in aggregate (6.65%/12.9% WER), which could be due to both the size of the dataset (only half the size of VoxCeleb2), or the makeup of the dataset (no speech-specific data).

All three pre-training datasets outperform from scratch training for ASR (even Kinetics), indicating that *pre-training on any amount or type of audio-visual data can be helpful*. We note that while Kinetics has the worst overall performance, it improves the most with mid-training (Rel. WER improvement of 14.03%) vs VoxCeleb2 (9.45%) and LRS-3 (6.17%) (Figure 3). These results confirm that the model pre-trained on Kinetics has the most to gain from language-representation alignment (as it contains no speech data), and training on LRS-3, which consists of primarily clean data, has less to gain.

The best ASR results with MAE and CLR are obtained on the LRS-3 pre-training dataset. However the best MAE+CLR performance was in using the Kinetics dataset. While it can be difficult to disentangle the results from pre-training

Method	PT	MT	en-de ↓		en-it ↓		en-nl ↓	
			Test-clean	Test-other	Test-clean	Test-other	Test-clean	Test-other
No Pre-training	None	-	6.84 ± 0.22	12.91 ± 0.47	6.84	12.91	6.84	12.91
MAE	K600	-	7.54	13.88	7.54	13.88	7.54	13.88
	K600	✓	5.69	11.34	5.95	12.55	5.73	12.04
	VC2	-	5.28	11.51	5.28	11.51	5.28	11.51
	VC2	✓	5.11	11.12	5.56	10.42	5.64	12.46
	LRS3	-	4.73	10.27	4.73	10.27	4.73	10.27
	LRS3	✓	5.61	10.85	**4.21**	**9.53**	5.32	10.33
CLR	K600	-	6.85	12.92	6.85	12.92	6.85	12.92
	K600	✓	5.02	10.85	4.72	10.62	4.65	10.41
	VC2	-	6.47	12.42	6.47	12.42	6.47	12.42
	VC2	✓	6.43	12.31	5.1	10.61	4.62	10.77
	LRS3	-	6.35	12.12	6.35	12.12	6.35	12.12
	LRS3	✓	6.74	10.59	5.84	11.33	6.01	10.13
MAE + CLR	K600	-	5.56	11.91	5.56	11.91	5.56	11.91
	K600	✓	5.02	11.68	5.23	11.37	6.39	12.03
	VC2	-	6.75	12.11	6.75	12.11	6.75	12.11
	VC2	✓	5.36	11.22	4.77	10.84	5.03	10.73
	LRS3	-	7.51	12.54	7.51	12.54	7.51	12.54
	LRS3	✓	7.16	12.29	5.08	11.13	6.17	12.32

Table 1: Performance (WER) on the Librispeech test-clean and test-other datasets with and without mid-training, and across Kinetics (K600), Voxceleb2 (VC2) and LRS-3 pre-training datasets. MT: With mid-training. MAE: Masked Autoencoding, CLR: Contrastive Learning. PT: Pre-Training. Underline denote the consistent WER drop through mid-training alone across the 3 datasets and PT strategies. We observe that translation Mid-training task benefit the global representations of CLR more consistently compared to MAE. Overall, it improves the 'only pre-trained' performance by aligning the learnt features towards the downstream task through auxiliary translation task. Further, italian language is the most effective as a mid-training task, suggesting that the languages that are complimentary to English may be more useful than others.

dataset size, this result may suggest that multi-task learning is more effective on out-of-domain data, where modalities contain non-redundant audio information, compared to VoxCeleb/LRS-3, where modalities consist of primarily redundant information.

MAE outperforms CLR, MAE+CLR on ASR: For ASR results averaged over all pre-training datasets, we find that MAE (5.63%/11.53% WER) alone outperforms both CLR (6.00%/11.67%) and MAE+CLR (6.09%/11.85%), suggesting that pre-training with masked auto-encoding objectives remains a promising approach for future exploration. Following intuition from Chan et al. (2022), it is likely that CLR-augmented methods outperform on more global downstream tasks, whereas MAE encodes more local information which is useful for ASR, and MAE+CLR is a useful mix of both. This hypothesis is validated in our experiments on SUPERB (Yang et al., 2021), where we found MAE+CLR most effective when aggregated across the mix of global (Intent Classification, Keyword Spotting), and local (Phoneme Recognition) tasks.

Mid-Training is most effective with multi-task pre-training: We explore the performance of our methods on four tasks from the SUPERB (Yang et al., 2021) benchmark in Figure 2. For SUPERB, mid-training improves performance for MAE+CLR models across most tasks. The notable exception is speaker diarization (SD), where there is minimal task overlap between SD and the mid-training target. Intent Classification (IC) is most improved (results not show in the tables), primarily due to a improvements in models pre-trained on

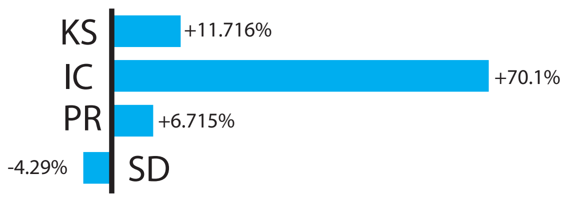

Figure 2: Aggregate (dataset/language) relative performance improvement (higher is better) under mid-training for MAE + CLR on SUPERB. KS: Keyword spotting, IC: Intent Classification, PR: Phoneme Recognition, SD: Speaker Diarization. We observe consistent improvement in performance due to translation mid-training on tasks which require local feature information (KS, IC and PR) whereas global task SD observe a decrease in performance. It further shows that translation mid-training task enhances the pre-trained model's performance for local feature tasks while hurts the global feature task.

the Kinetics (+80.17%) and LRS-3 (+102.30%) datasets, which benefit from the additional textual alignment. Keyword spotting (KS) improvements can also be largely attributed improvements on models pre-trained on Kinetics (+27.52%), for similar reasons. Models pre-trained on VoxCeleb2 improve less with mid-training compared to models pre-trained with both Kinetics and LRS-3 for all tasks. We posit that since VoxCeleb2 dataset is already multi-lingual, and benefits less from further multi-lingual training.

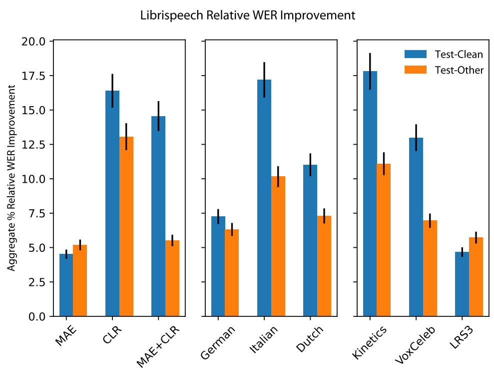

Figure 3: Average relative WER improvement on the Librispeech test-clean and test-other datasets with mid-training to show the effect of pre-training methods (left), mid-training translation pairs (center), and pre-training datasets (right). Translation mid-training improves upon CLR pre-training the most as it aligns its features for the local information required for ASR. Among the translation languages, Italian provides the best improvement, suggesting a complimentary language to English gains the most compared to languages that shares its roots with English. Models pre-trained on non-speech dataset Kinetics benefit the most from translation mid-training followed by noisy speech dataset Voxceleb2 and then clean speech dataset LRS3.

Note on baseline conformer performance: In this work, we note that our baseline conformer models do not match the performance of Gulati et al. (2020). Note that our primary goals was not to attain state of the art models, but study the impact of pre-training methods and datasets on ASR performance. The higher WER can be attributed to lower batch size used in out experiments, which was done to account for the large number of ablation studies done for this paper. While the overall baseline performance may be worse, the insights learned from the relative performance comparisons across the large-scale ablation are transferable to larger, more expensive models.

In summary, our results indicate the following:

- Audio-visual pre-training is effective in almost all scenarios.
- Mid-training is useful and including data which is complementary is more effective than including data similar to pre-training data.
- Clean speech audio-visual dataset LRS-3 is an effective pre-training dataset given its size, compared to Kinetics and Voxceleb2.
- MAE pre-training is more effective than contrastive learning in ASR, while augmenting pre-training with CLR can help with downstream tasks that use global information.

5. Discussion

Recently, the size of pre-trained models and the datasets have increased to such an extent that it is cost-prohibitive to pre-train these models on datasets aligned with the downstream tasks of interest. Hence, a light-weight mid-training strategy can tune the pre-trained features strengthening the downstream performance.

An alternative to the mid-training strategy is to include task during pre-training itself. This alternative strategy has two drawbacks; first, the amount of labeled data available for the mid-training task is typically not large enough to have significant impact when jointly learned in the pre-training stage. Secondly, the mid-training approach is more practical as it can be applied to already available pre-trained models instead of training the models from scratch which requires large amounts of time and compute.

6. Conclusion & Future Directions

This work presents a multi-lingual mid-training objective and a large-scale analysis of multiple audio-visual pre-training methods and datasets, which confirms observations from (Hsu et al., 2021) and (Chan et al., 2022) — we show how large scale audio-visual pre-training significantly improves downstream ASR performance, and that a well-chosen mid-training task can help the final downstream task.

While this paper presents initial insights into how mid-training tasks impact models multi-modal pre-trained models, we believe that significant additional future work remains to fully understand how sequences of training tasks can align large pre-trained models with downstream tasks.

One interesting direction for future work is an exploration of additional mid-training tasks. In this work, we show that translation has the power to bridge gaps between multi-modal pre-trained models and language-based ASR tasks. Paired data for translation data can often be scarce, and may not be the optimal choice for future mid-training tasks. Instead, it may be insightful to explore mid-training tasks which are centered around synthetic data (such as TTS data from text datasets, or text generated by large language models) or self-supervised approaches to mid-training.

Another closely related direction of future work explores how pre-training tasks impact the performance of downstream and mid-trained models. Here, we focus on multi-modal pre-training, as it is a key emergent direction of ASR research. However mid-training can easily be applied to uni-modal pre-training, or even zero-shot transfer from foundational models.

In conclusion, this study sheds light on the impact of mid-training tasks in the context of multi-modal pre-training and demonstrates the significant improvement in downstream automatic speech recognition performance achieved through large-scale audio-visual pre-training. By continuing to delve into these areas, we can advance our understanding of how to effectively align pre-trained models with diverse downstream tasks and unlock new possibilities for multi-modal ASR research.

7. Bibliographical References

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. 2016. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)*, pages 265–283.

Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. 2018. Lrs3-ted: a large-scale dataset for visual speech recognition. *arXiv:1809.00496*.

Hassan Akbari, Linagzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. 2021. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *arXiv:2104.11178*.

Jean-Baptiste Alayrac, Adria Recasens, Rosalia Schneider, Relja Arandjelovic, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. 2020. Self-supervised multimodal versatile networks. *NeurIPS*, 2(6):7.

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. 2021. Vivit: A video vision transformer. *arXiv:2103.15691*.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv:1607.06450*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. Wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460.

Junwen Bai, Bo Li, Yu Zhang, Ankur Bapna, Nikhil Siddhartha, Khe Chai Sim, and Tara N Sainath. 2022. Joint unsupervised and supervised training for multilingual asr. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6402–6406. IEEE.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. 2021. Is space-time attention all you need for video understanding? *arXiv:2102.05095*.

Alexander Bukharin and Tuo Zhao. 2023. Data diversity matters for robust instruction tuning. *arXiv preprint arXiv:2311.14736*.

Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. 2018. A short note about kinetics-600. *arXiv:1808.01340*.

Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308.

David M. Chan and Shalini Ghosh. 2022. Content-context factorized representations for automated speech recognition.

David M Chan, Shalini Ghosh, Debmalya Chakrabarty, and Björn Hoffmeister. 2022. Multimodal pre-training for automated speech recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 246–250. IEEE.

David M. Chan, Shalini Ghosh, Ariya Rastrow, and Björn Hoffmeister. 2023. Using external off-policy speech-to-text mappings in contextual end-to-end automated speech recognition.

William Chan, Daniel Park, Chris Lee, Yu Zhang, Quoc Le, and Mohammad Norouzi. 2021. Speechstew: Simply mix all available speech recognition data to train one large neural network. *arXiv:2104.02133*.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Chung-Cheng Chiu, Arun Narayanan, Wei Han, Rohit Prabhavalkar, Yu Zhang, Navdeep Jaitly, Ruoming Pang, Tara N Sainath, Patrick Nguyen, Liangliang Cao, et al. 2021. Rnn-t models fail to generalize to out-of-domain audio: Causes and solutions. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 873–880. IEEE.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv:2204.02311*.

Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. 2018. Voxceleb2: Deep speaker recognition. *arXiv:1806.05622*.

Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu. 2021. W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 244–250. IEEE.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805*.

Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. MuST-C: a Multilingual Speech Translation Corpus. In *NAACL 2019*, pages 2012–2017.

Alexey Dosovitskiy, Lucas Beyer, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv:2010.11929*.

Christoph Feichtenhofer, Haoqi Fan, Yanghao Li, and Kaiming He. 2022. Masked autoencoders as spatiotemporal learners. *arXiv:2205.09113*.

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211.

Shalini Ghosh, Patrick Lincoln, Ashish Tiwari, and Xiaojin Zhu. 2016. Trusted machine learning for probabilistic models. *ICML Workshop on Reliable Machine Learning in the Wild*.

Priya Goyal, Quentin Duval, Isaac Seessel, Mathilde Caron, Ishan Misra, Levent Sagun, Armand Joulin, and Piotr Bojanowski. 2022. Vision models are more robust and fair when pretrained on uncurated images without supervision. *arXiv preprint arXiv:2202.08360*.

Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. 2013. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6645–6649. Ieee.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. 2020. Conformer: Convolution-augmented transformer for speech recognition. *arXiv:2005.08100*.

Wei Han, Zhengdong Zhang, Yu Zhang, Jiahui Yu, Chung-Cheng Chiu, James Qin, Anmol Gulati, Ruoming Pang, and Yonghui Wu. 2020. Contextnet: Improving convolutional neural networks for automatic speech recognition with global context. *arXiv:2005.03191*.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. *arXiv:2203.15556*.

Wei-Ning Hsu and Bowen Shi. 2022. U-hubert: Unified mixed-modal speech pretraining and zero-shot transfer to unlabeled modality. In *Advances in Neural Information Processing Systems*.

Wei-Ning Hsu, Yao-Hung Hubert Tsai, Benjamin Bolte, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: How much can a bad teacher benefit asr pre-training? In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6533–6537. IEEE.

Ronghang Hu and Amanpreet Singh. 2021. Unit: Multimodal multitask learning with a unified transformer. *arXiv:2102.10772*.

Yash Jain, Harkirat Behl, Zsolt Kira, and Vibhav Vineet. 2024. Damex: Dataset-aware mixture-of-experts for visual understanding of mixture-of-datasets. *Advances in Neural Information Processing Systems*, 36.

Mostafa Karimi, Changliang Liu, Kenichi Kumatani, Yao Qian, Tianyu Wu, and Jian Wu. 2022. Deploying self-supervised learning in the wild for hybrid automatic speech recognition. *arXiv:2205.08598*.

Shigeki Karita, Nanxin Chen, Tomoki Hayashi, Takaaki Hori, Hirofumi Inaguma, Ziyan Jiang, Masao Someki, Nelson Enrique Yalta Soplin, Ryuichi Yamamoto, Xiaofei Wang, et al. 2019. A comparative study on transformer vs rnn in speech applications. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 449–456. IEEE.

Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, et al. 2024. Droid: A large-scale in-the-wild robot manipulation dataset. *arXiv preprint arXiv:2403.12945*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv:1412.6980*.

Sachit Kuhar, Shuo Cheng, Shivang Chopra, Matthew Bronars, and Danfei Xu. 2023. Learning to discern: Imitating heterogeneous human demonstrations with preference and representation learning. In *Conference on Robot Learning*, pages 1437–1449. PMLR.

Rimita Lahiri, Kenichi Kumatani, Eric Sun, and Yao Qian. 2021. Multilingual speech recognition using knowledge transfer across learning processes. *arXiv:2110.07909*.

Jinyu Li, Rui Zhao, Zhong Meng, Yanqing Liu, Wenning Wei, Sarangarajan Parthasarathy, Vadim Mazalov, Zhenghao Wang, Lei He, Sheng Zhao, et al. 2020. Developing rnn-t models surpassing high-performance hybrid models with customization capability. *arXiv:2007.15188*.

Vladislav Lialin, Stephen Rawls, David Chan, Shalini Ghosh, Anna Rumshisky, and Wael Hamza. 2023. Scalable and accurate self-supervised multimodal representation learning without aligned video and text data. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*.

Tatiana Likhomanenko, Qiantong Xu, Vineel Pratap, Paden Tomasello, Jacob Kahn, Gilad Avidov, Ronan Collobert, and Gabriel Synnaeve. 2020. Rethinking evaluation in asr: Are our models robust enough? *arXiv:2010.11745*.

Sridhar Mahadevan, Bamdev Mishra, and Shalini Ghosh. 2018. A unified framework for domain adaptation using metric learning on manifolds. *CoRR*, abs/1804.10834.

Ajay Mandlekar, Danfei Xu, Josiah Wong, Soroush Nasiriany, Chen Wang, Rohun Kulkarni, Li Fei-Fei, Silvio Savarese, Yuke Zhu, and Roberto Martín-Martín. 2022. What matters in learning from offline human demonstrations for robot manipulation. In *Conference on Robot Learning*, pages 1678–1690. PMLR.

Vimal Manohar, Pegah Ghahremani, Daniel Povey, and Sanjeev Khudanpur. 2018. A teacher-student learning approach for unsupervised domain adaptation of sequence-trained asr models. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 250–257. IEEE.

Soumyajit Mitra, Swayambhu Nath Ray, Bharat Padi, Arunasish Sen, Raghavendra Bilgi, Harish Arsikere, Shalini Ghosh, Ajay Srinivasamurthy, and Sri Garimella. 2022. Unified modeling of multi-domain multi-device asr systems.

Ladislav Mošner, Minhua Wu, Anirudh Raju, Sree Hari Krishnan Parthasarathi, Kenichi Kumatani, Shiva Sundaram, Roland Maas, and Björn Hoffmeister. 2019. Improving noise robustness of automatic speech recognition via parallel data and teacher-student learning. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6475–6479. IEEE.

Arun Narayanan, Ananya Misra, Khe Chai Sim, Golan Pundak, Anshuman Tripathi, Mohamed Elfeky, Parisa Haghani, Trevor Strohman, and Michiel Bacchiani. 2018. Toward domain-invariant speech recognition via large scale training. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 441–447. IEEE.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv:1807.03748*.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. 2019. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv:1904.08779*.

Santiago Pascual, Mirco Ravanelli, Joan Serra, Antonio Bonafonte, and Yoshua Bengio. 2019.

Learning problem-agnostic speech representations from multiple self-supervised tasks. *arXiv:1904.03416*.

Tivadar Pápai, Shalini Ghosh, and Henry Kautz. 2012. Combining subjective probabilities and data in training Markov Logic Networks. volume 7523, pages 90–105.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv:2212.04356*.

Vivek Ramanujan, Thao Nguyen, Sewoong Oh, Ali Farhadi, and Ludwig Schmidt. 2023. On the connection between pre-training data diversity and fine-tuning robustness. In *Advances in Neural Information Processing Systems*, volume 36, pages 66426–66437. Curran Associates, Inc.

Aaqib Saeed, David Grangier, and Neil Zeghidour. 2021. Contrastive learning of general-purpose audio representations. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3875–3879. IEEE.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. Wav2vec: Unsupervised pre-training for speech recognition. *arXiv:1904.05862*.

Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. 2022. Learning audio-visual speech representation by masked multimodal cluster prediction. *arXiv:2201.02184*.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7464–7473.

Gabriel Synnaeve, Qiantong Xu, Jacob Kahn, Tatiana Likhomanenko, Edouard Grave, Vineel Pratap, Anuroop Sriram, Vitaliy Liptchinsky, and Ronan Collobert. 2019. End-to-end asr: from supervised to semi-supervised learning with modern architectures. *arXiv:1911.08460*.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9.

Chaitanya Talnikar, Tatiana Likhomanenko, Ronan Collobert, and Gabriel Synnaeve. 2021. Joint masked cpc and ctc training for asr. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3045–3049. IEEE.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. 2022. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *arXiv:2203.12602*.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2019, page 6558. NIH Public Access.

Lifu Tu, Garima Lalwani, Spandana Gella, and He He. 2020. An empirical study on robustness to spurious correlations using pre-trained language models. *Transactions of the Association for Computational Linguistics*, 8:621–633.

Kostiantyn Tyshchenko et al. 2000. Metatheory of linguistics. *Osnovy*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Chengyi Wang, Yu Wu, Yao Qian, Kenichi Kumatani, Shujie Liu, Furu Wei, Michael Zeng, and Xuedong Huang. 2021a. Unispeech: Unified speech representation learning with labeled and unlabeled data. In *International Conference on Machine Learning*, pages 10937–10947. PMLR.

Luyu Wang, Pauline Luc, Adria Recasens, Jean-Baptiste Alayrac, and Aaron van den Oord. 2021b. Multimodal self-supervised learning of general audio representations. *arXiv:2104.12807*.

Shinji Watanabe, Takaaki Hori, Jonathan Le Roux, and John R Hershey. 2017. Student-teacher network learning with enhanced features. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5275–5279. IEEE.

Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. 2021. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv:2109.14084*.

Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y Lin, Andy T Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, et al. 2021. Superb: Speech processing universal performance benchmark. *arXiv:2105.01051*.

Jie Zhang, Junting Zhang, Shalini Ghosh, Dawei Li, Jingwen Zhu, Heming Zhang, and Yalin Wang. 2020a. Regularize, expand and compress: Non-expansive continual learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*.

Qian Zhang, Han Lu, Hasim Sak, Anshuman Tripathi, Erik McDermott, Stephen Koo, and Shankar Kumar. 2020b. Transformer transducer: A streamable speech recognition model with transformer encoders and rnn-t loss. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7829–7833. IEEE.

Yu Zhang, Wei Han, James Qin, Yongqiang Wang, Ankur Bapna, Zhehuai Chen, Nanxin Chen, Bo Li, Vera Axelrod, Gary Wang, et al. 2023. Google usm: Scaling automatic speech recognition beyond 100 languages. *arXiv preprint arXiv:2303.01037*.

Zi-qiang Zhang, Yan Song, Jian-shu Zhang, Ian Vince McLoughlin, and Li-rong Dai. 2020c. Semi-supervised end-to-end asr via teacher-student learning with conditional posterior distribution. In *INTERSPEECH*, pages 3580–3584.

Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. 2022. Simple multi-dataset detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7571–7580.